ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

When “Used For” is not Equivalent: Exploring the Potential of Large Language Models for Vocabulary Relationship Analysis and Standardization

Andreas Oskar Kempf, Arben Hajra

ZBW – Leibniz Information Centre for Economics, Germany

Tobias Potoczki

Catholic University of Eichstätt-Ingolstadt, Germany

Abstract

An equivalence relationship in a thesaurus may include quasi-synonyms and even antonyms. At a time when neuro-symbolic AI approaches are gaining momentum and knowledge organization systems (KOS) are increasingly transforming into knowledge bases - for instance, serving as sources in retrieval-augmented generation (RAG) systems - this structural characteristic appears no longer well-suited to emerging application scenarios. Instead, there is a need for greater semantic standardization of authority data regarding synonym control. At the same time, clear semantic boundaries of a concept increase its interoperability. Using the domain-specific, bilingual STW Thesaurus for Economics as an example, this study examines the potential of state-of-the-art large language models (LLMs) to analyze vocabulary relationships as a preliminary step toward concept standardization. Focusing on the relationship between a preferred term and its non-preferred terms, we test how semantically precise LLMs can describe these relationships. The evaluation is based on an intellectually created benchmark and an optimized prompt strategy. Our methodology combines two approaches: first, we evaluate LLM performance using standard metrics and pre-defined relationship types. Second, we combine the best-performing models in a cascading architecture based on self-reporting confidence scores. The results provide insights into the extent to which off-the-shelf LLMs can assist in decision-making for concept splitting.

1.0 Introduction

In a thesaurus, equivalence relationships can also be established between quasi-synonyms. Depending on the required degree of differentiation and the subject area of the thesaurus, the relationship between the preferred term and its corresponding non-preferred term(s) might also be generic, partitive, or instantial. LLMs are considered highly suitable for uncovering the exact relationship between terms. Trained on extensive corpora, they can capture rich patterns of synonyms, contextual variations, and domain-specific usage. Through contextual embeddings and next-token prediction, they identify detailed semantic differences beyond simple surface-level similarity. Unlike the traditional labor-intensive

manual maintenance of thesauri, LLMs can dynamically evaluate term relationships in context, assess whether a non-preferred term reflects a hierarchical or associative relationship, and generate explanations for these decisions. These capabilities could support thesaurus modeling workflows by reducing manual effort and improving scalability. This study therefore addresses the following research questions:

- To what extent can LLMs assist in determining the semantic relationship between a preferred term and its non-preferred term in a thesaurus?
- To what extent could this relationship assessment be reliably used as a decision-making tool and serve as a foundation for human-LLM collaboration?

To answer these questions, the study pursues two objectives: First, it systematically evaluates the ability of state-of-the-art LLMs to classify semantic relationships between preferred and non-preferred terms in a domain-specific thesaurus. Second, it investigates how such classifications can support editorial decision-making in concept refinement and potential splits. Methodologically, a ground truth (GT) dataset based on expert evaluation is compared with LLM-generated classifications, using both individual model assessments and an ensemble approach. Performance is evaluated using standard metrics and agreement analyses.

The paper is structured as follows: Section 2 reviews theoretical foundations and related work. Section 3 introduces the data source, sampling strategy, classification schema, and LLM setup. Section 4 describes the research procedure and evaluation framework. Section 5 presents the results, and Section 6 discusses their implications for human–LLM collaboration and provides an outlook.

2.0 Background and Related Work

2.1 The Equivalence Relationship in a Thesaurus

According to ISO 25964-1 (2011), the equivalence relationship is “the relationship between a preferred term and its corresponding non-preferred term(s)” (44). It applies between terms rather than between concepts. The standard includes cases in which related concepts are treated as quasi-synonyms, as well as specific terms that have a generic, partitive, or instance relationship to the preferred term and are subsumed under a broader concept. In a monolingual thesaurus, this relationship exists between terms in the same language, while between different languages it is referred to as cross-language equivalence.

This approach reflects the idea that “concepts are so closely related that discussion of one is likely to interest a user seeking the other” (48). The extent to which terms are treated as quasi-synonyms would depend on the subject field and the level of discrimination required during searching. Subsuming specific terms under a broader concept could reduce the number of preferred terms in a thesaurus. However, this limits the specificity of indexing and searching and is mainly beneficial when collections contain little information on a given subject. In such cases, non-preferred terms provide additional entry points to support user access (48).

However, the environment and application contexts of thesauri have changed significantly in recent years. They are now often integrated into search systems that combine multiple databases with different indexing vocabularies. ISO 25964-2 (2013) addresses this issue by focusing on interoperability with other vocabularies. At the same time, automated subject indexing and modern embedding models require semantically well-defined concepts. This also applies to vector-based representations of thesaurus concepts, for example, in RAG systems. In this context, non-preferred terms that are actually hierarchically or otherwise semantically related to the preferred term can become problematic. Such terms, therefore, need to be identified as potential candidates for distinct future concepts.

2.2 Quality Assessment in Synonym Control

The critical review of synonym control is described in the ISO standard by the change type "j) splitting a concept into two or more, usually selecting an existing non-preferred term to be a preferred term" (ISO 25964-1 2011, 96). The review procedure requires editors to monitor the frequency of preferred term usage and identify highly used terms as candidates for revision while maintaining the consistency of the overall structure. In addition, integrated search environments increase the importance of interoperability, making it necessary to examine how topics are represented in related vocabularies. Overall, updating synonym control is time-consuming and strongly dependent on human expertise.

These factors highlight the need for robust and scalable approaches. To investigate the potential of LLMs for specifying the relationship between preferred and non-preferred terms, we apply sampling parameters that simulate the manual processing steps described above. This ensures that the dataset includes descriptor records with numerous non-preferred terms that are not strictly equivalent. The parameters include the number of non-preferred terms, the frequency of concept usage, and the number of non-exact mapping relationships to concepts in other vocabularies.

2.3 LLMs for Vocabulary Relationship Analysis

Recent research shows a growing interest in the application of LLMs for important knowledge organization tasks, particularly for identifying and classifying semantic relationships between terms (Aggarwal et al. 2026). Studies in ontology and thesaurus engineering indicate that, when guided by explicit relationship schemas, LLMs can classify relations such as synonymy and hierarchical (broader–narrower) relations with promising accuracy, as shown in benchmark evaluations and ontology generation experiments (Li et al. 2026; Aggarwal et al. 2026). Prompt-based approaches using generative LLMs have also been found to rival—and sometimes even outperform—traditional ontology matching systems, especially in identifying equivalence or close-match relationships while requiring little task-specific training data (He et al. 2023). Beyond relationship classification, LLMs are increasingly explored as assisting tools for ontology and vocabulary development, for example, to identify missing concepts, detect potential misclassifications, or highlight insufficiently specified definitions (Babaei Giglou et al. 2023).

Although performance varies across domains and depends on prompt design, model selection, and evaluation setup, the literature generally suggests that LLMs are most effective as prioritization and decision-support tools within human-in-the-loop workflows rather than as fully autonomous ontology or thesaurus curators (Shimizu and Hitzler 2025; S. Chen 2025). Focusing on meronomy, Proietti et al. (2025) examine whether LLMs capture the logical properties of part-whole relations and find only partial semantic competence when relational constraints are explicitly tested. Cao et al. (2025) propose an evaluation framework covering multiple semantic relationships and show that while LLMs approximate relational patterns, they do not robustly assess structural constraints. Roussinov et al. (2023), studying classification tasks such as part-of or property-of relations, also identify shortcomings but conclude that LLM evaluations "come remarkably close to achieving human-level performance on rigorous benchmarks" (1463).

2.4 Prompt Framework and Prompt Strategy

Recent work on ontology and relation classification shows that LLM performance in identifying semantic relations largely depends on prompt design (S. Chen 2025; Aggarwal et al. 2026). Prompt engineering acts as a methodological control mechanism that reduces variability and activates task-specific reasoning patterns (B. Chen et al. 2025), which is particularly critical in vocabulary relationship analysis, where explicit instructions and output schemas ensure terminological consistency and limit ambiguity.

S. Chen (2025) proposes a six-element prompt framework - task specification, domain context, role definition, input data, procedural guidance, and output constraints - allowing systematic variation and iterative refinement to stabilize semantic interpretation and improve relational inference. Iterative calibration is emphasized in studies on advanced prompting techniques and relation classification (B. Chen et al. 2025; Aggarwal et al. 2026). In vocabulary analysis, linguistic sensitivity is a central aspect of the design. In multilingual settings, contextual enrichment and explicit semantic framing are necessary to disambiguate homonyms and near-synonyms across subdomains (S. Chen 2025).

3.0 Research Design

3.1 Data Source

The STW Thesaurus for Economics, available in English and German since 2007, standardized four controlled source vocabularies and includes many non-preferred or non-descriptors that are not exact synonyms (see Figure 1). Its seven sub-thesauri-V Economics, B Business economics, W Economic sectors, P Commodities, N Related subject areas, G Geographic names, and A General descriptor-contain nearly 6,200 descriptor records, over 7,500 non-descriptors in English, and nearly 14,800 in German. The structure is symmetrical. Descriptor records cover up to ten hierarchical levels, can include up to five different subject categories, and are linked to other vocabularies or authority data such as GND or Wikidata (Kempf and Neubert 2021).



Figure 1. Display of the STW descriptor record for “Participative leadership”

3.2 Sampling Parameters for Ground Truth

For each language, samples of 100 descriptors and their corresponding non-descriptors were drawn. The German set included 775 preferred/non-preferred term pairs, the English 468. Sampling followed six criteria:

1. language
2. sub-thesaurus affiliation: selection of 20 descriptor records from the five STW sub-thesauri V, B, W, P, and N; operationalized by clear assignment to one of these via corresponding subject categories; the descriptor records from sub-thesaurus P could also have been assigned to other sub-thesauri due to a lack of cases that met all six criteria
3. number of non-preferred terms per preferred term (in descending order)
4. hierarchical level of a concept within the overall thesaurus structure (≤5/10)
5. frequency of use (≥500; exception: records from sub-thesaurus P)
6. number of non-equivalence mappings to GND or Wikidata (≥3 in total)

3.3 Human Evaluation and Relationship Classification

The human evaluation was conducted by two domain experts from the thesaurus editorial board who first assessed cases outside the sample and discussed them to minimize inconsistency in evaluation between them. Evaluation relied on five relationship categories (Table 1), also forming the basis for LLM assessment.

Table 1 Relationship Classification Schema

Relationship Type	**Relationship Description (aligned with the prompt)**
Equivalent (0)	Relationship between descriptor (D) and non-descriptor (ND) corresponds to an equivalence relationship and D and ND are therefore synonymous
Hierarchical (1)	Relationship... corresponds to a generic or partitive relationship, and thus a hierarchical relationship
Associative (2)	Relationship... corresponds to an associative relationship: D and ND are conceptually related but neither synonymous nor linked through a hierarchical or instance relationship.
Instantial (3)	Relationship... corresponds to an instance relationship and ND represents a specific instance of the concept behind D, usually in form of a named entity or proper name
Other (4)	Relationship... does not correspond to any of the relationship types mentioned above

The samples revealed uneven distributions: around two-thirds of term pairs were hierarchical, and almost 15% associative. Exact equivalences accounted for 20% of German pairs and 12% of English pairs. In the German sample, “instance” or “other” relationships were rare. In the English sample, there were none of them.

4.0 Research Procedure

The study followed a structured workflow (Figure 2) to ensure comparability between expert assessments and LLM-based assessments:

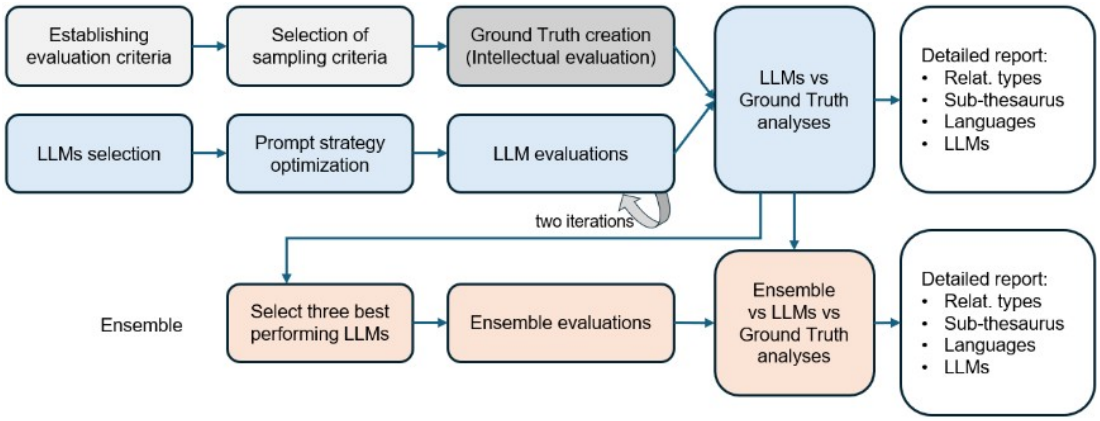

Figure 2. Flowchart of the whole research setup

After establishing the evaluation criteria and sampling parameters, the sample was drawn, and the GT was determined using the relationship classification scheme (Table 1). This dataset served as the reference for all subsequent analyses. In parallel, a set of LLM candidates was selected according to the criteria described in Section 4.1. Prompt design was treated as an iterative calibration process, with multiple prompt versions tested on a smaller subset of data to refine interpretation and optimize output structure. Each model was then applied to the full dataset to generate relationship classifications and explanations, with multiple runs performed for some models. The outputs were compared to the GT using metrics such as relationship type, sub-thesaurus, language, and model.

The ensemble approach employs a cascading architecture that incorporates self-reported confidence scores. Each LLM provides both a classification decision and a confidence score, which are weighted by model-specific performance coefficients to calculate the system's overall confidence for each prediction. The cascade begins with three lightweight, high-performing models. If their agreement exceeds a predefined threshold (>85%), classification is finalized; otherwise, additional models are invoked sequentially-a medium-performance model first, followed by an arbitration model if disagreement persists. This approach produces both instance-level confidence scores and dataset-level quality metrics, supporting editorial decisions and enabling tiered automation within the thesaurus maintenance workflow. The ensemble's performance was evaluated using the same metrics as the individual models to allow for direct comparison.

4.1 LLM Implementation

The test setup included ten LLMs from six families (GPT-OSS, Qwen, DeepSeek, Llama, Mistral, and Apertus). Model selection was based on three criteria: (1) scale (number of parameters), (2) accessibility (preference for open-weight or open-source model), and (3) practical suitability, including proven multilingual capabilities for English and German. Implemented models were Apertus (8B, 70B Instruct), DeepSeek-R1 (8B, 70B), GPT-OSS (20B (*), 120B (*)), Mistral (8B), Llama 3 (8B), Llama 3.3 (70B Instruct), Qwen3 (32B (*)).

All models were executed locally via Ollama on a server with 2 TB RAM and two NVIDIA Quadro RTX 8000 GPUs (64 GB VRAM each). Internet access, web search, retrieval-augmented generation (RAG), external databases, and other external knowledge sources were disabled during all experiments. Consequently, relationship classifications were generated solely from the information contained in the prompt and the models' pretrained parametric knowledge.

In the first evaluation phase (Figure 2), all models were tested on the dataset. The three best-performing models (GPT-OSS 20B, GPT-OSS 120B, and Qwen3 32B, marked with *) underwent a second iteration to assess the stability of the results.

4.2 Optimized Prompt Strategy and Evaluation Management System

The optimized prompt strategy consisted of six components, detailed in Appendix Table A1. Beyond the core relationship classification task, the prompt instructed the models to generate a one-sentence definition for each descriptor. These definitions served as supplementary outputs for documentation and transparency purposes only; they were not used as inputs or intermediate representations at any stage of the classification process. Instead, each relationship assessment was derived directly from the descriptor, its corresponding non-descriptor, the STW subject-category context, and the relationship classification schema presented in Table 1.

A dedicated interface was developed to manage human and LLM assessment and generate evaluation reports, including an overall performance summary, comparative performance metrics, performance by sub-thesaurus, confusion matrices, per-class performance metrics, disagreement patterns, and detailed disagreement cases for each LLM (for example, Appendix Figures B1 and B2).

5.0 Results

We present the results according to the two approaches we have chosen.

5.1 Single-Model Performance

In the German sample, GPT-OSS 20B achieved the highest overall accuracy with an F1 score of 70.22% and the lowest disagreement rate (29.8%). The most frequent mismatch (73 cases) occurred between GT "equivalence" ratings and model "hierarchical" classifications. Performance varied across relationship types and metrics, with one of the GPT-OSS iterations generally performing best:

- Equivalent (0): Llama 3.3 70B detected most cases (129); GPT-OSS 20B had the highest F1 (66.2%).
- Hierarchical (1): GPT-OSS 20B performed best (413 cases, Table 2a, F1 79.6%).

- Associative (2): DeepSeek-R1 70B detected the most cases (55); GPT-OSS 120B achieved the highest F1 (52.2%).
- Instance (3): Qwen3 32B detected most cases (11); Llama 3.3 70B had the highest F1 (63.2%).
- Other (4): GPT-OSS 20B performed best, though only a few cases existed.

Table 2 Confusion matrices for German and English datasets, illustrating model vs. GT prediction across the five relationship types for the best overall performing LLMs.

a. GPT-OSS 20B (German)

GT \ LLM	0	1	2	3	4
Equivalence (0)	**77**	73	3	3	1
Hierarchical (1)	23	**413**	19	28	5
Associative (2)	2	61	**41**	1	5
Instance (3)	0	2	0	**9**	0
Other (4)	1	1	1	0	**0**

a. GPT-OSS 20B (English)

GT \ LLM	0	1	2	3	4
Equivalence (0)	**50**	26	4	1	0
Hierarchical (1)	12	**257**	29	22	0
Associative (2)	2	27	**31**	6	1
Instance (3)	0	0	0	**0**	0
Other (4)	0	0	0	0	**0**

At the individual term level, inconsistencies occurred due to spelling differences in non-preferred terms. For example, in the descriptor record "*Liquidation*", GPT-OSS 20B iteration 1 classified "*Unternehmensstilllegung*" as hierarchical but "*Unternehmensstillegung*" as equivalent; iteration 2 classified both as hierarchical.

For the English dataset, GPT-OSS 120B achieved the highest accuracy (F1 72.22%) and lowest disagreement rate (27.8%). The most frequent mismatch (29 cases) was between GT hierarchical and model associative ratings: Detailed performance by relationship type:

- Equivalent (0): Llama 3 8B detected most cases (72); Qwen3 32B highest F1 (73.1%).
- Hierarchical (1): GPT-OSS 120B performed best (257 cases, Table 2b, F1 81.6%).
- Associative (2): Qwen3 32B with highest performance (39 cases, F1 51.0%).

At the individual term level, cross-linguistic inconsistencies were identified. For example, for the descriptor record "Dictatorship" in English, "Military rule" was classified as hierarchical, while the German equivalent "*Militärherrschaft*" was classified as an instance.

Sub-thesaurus P (Commodities) nearly consistently showed the highest accuracy (up to 79.2% English and up to 78.8% German), whereas sub-thesaurus V (Economics) performed weakest (down to 49.6% English and 36.6% German).

5.2 Ensemble Approach

We implemented a cascading multi-model ensemble structured in three levels. Fast, lightweight models (GPT-OSS 20B, Qwen3 32B, DeepSeek-R1 8B) were applied first. If their predictions did not reach sufficient agreement, a medium-capacity model (Llama 3.3 70B) was added, and high-capacity models (GPT-OSS 120B) were reserved for final arbitration. At each stage, agreement was measured as the proportion of weighted votes supporting the leading classification class. The process ended early if agreement reached 85% or higher. If agreement fell below 70%, the medium model was engaged, and if the leading class still accounted for less than 60% of votes, the high-capacity arbitrator reviewed the reasoning from previous models.

Model weights were adjusted based on size and prior single-model performance. Because Associative relations were the most error-prone category, all such predictions underwent a second-pass verification, where the model was explicitly prompted to rule out a hierarchical relationship before confirming the assignment. Each output included a confidence score, a disagreement severity rating, and a flag for recommended human review.

The ensemble approach achieved 75.6% accuracy for English and 74.9% for German, surpassing the best individual models (72.2% and 70.2%, respectively; see Figure B2). Hierarchical relations were classified most reliably (F1: 83.3% / 82.3%), followed by equivalence (74% / 68.0%), while associative relations remained the most challenging (F1: 51.8% / 59.3%). Again, sub-thesaurus P (Commodities) consistently showed the highest accuracy (86.1% English, 88.3% German), whereas sub-thesaurus V (Economics) performed weakest (61.4% / 69.0%).

Confidence scores not only indicated reliability but also supported selective automation. In the English dataset, predictions with confidence 0.50-0.59 achieved 82.8% accuracy, those above 0.90 reached 91.5%, and predictions below 0.50 fell to 53.0%. High-confidence equivalence and hierarchical predictions were particularly precise (85.7% and 83.8%), while associative predictions remained unreliable (41.4%). These confidence scores enable a graded workflow, in which high-confidence classifications can be accepted automatically, and uncertain cases - especially associative predictions - are flagged for human review. Thresholds can be adjusted according to the desired quality requirements.

6.0 Discussion and Outlook

This study demonstrates that LLMs, when guided by optimized prompts and combined in a well-chosen ensemble, can effectively support human-LLM collaboration in standardizing

vocabulary relationships. LLMs reliably identify non-preferred terms that are not equivalent, and their explanations often suggest the most accurate relationship type. However, expert validation remains essential to ensure accuracy and precision. Optimized prompts and human oversight are therefore critical to improving editorial efficiency and scalability. Beyond editorial workflows, these findings are relevant for transforming traditional thesauri into machine-interpretable knowledge bases, where ambiguous equivalence relationships can hinder interoperability and automated knowledge integration. Automated refinement of KOS can ultimately enhance information retrieval in discovery and RAG systems that rely on KOS-based metadata.

Limitations of the study include uneven distribution of relationship types in the English sample, which lacked “instance” and “other” examples. Additionally, having more than two evaluators for the GT would have been advantageous, though both were experts from the thesaurus editorial board.

Future work will explore few-shot and chain-of-thought prompting strategies to improve agreement between LLMs and domain experts. We also plan to implement more efficient filtering for split candidate selection and to develop an interactive LLM-assisted KOS editing platform that combines multiple sub-scores to streamline editorial workflows. In addition, future research should investigate whether providing authoritative economics-specific definitions or scope notes derived from external domain knowledge bases can further improve relationship classification accuracy, consistency, and reproducibility compared to relying solely on the models’ pretrained knowledge. Overall, our results indicate that LLMs can play a meaningful supporting role in thesaurus maintenance and vocabulary standardization. When embedded within carefully designed human-in-the-loop workflows, they offer a promising pathway toward more scalable and semantically consistent KOS.

Acknowledgement

The authors would like to thank Manfred Faden, a member of the STW editorial board, for his support in creating the Ground Truth.

AI usage statement: The authors did not use generative AI tools or AI-assisted technologies in the creation, drafting, or editing of this document.

References

Aggarwal, Tanay, Angelo Salatino, Francesco Osborne, and Enrico Motta. 2026. “Large Language Models for Scholarly Ontology Generation: An Extensive Analysis in the Engineering Field.” Information Processing & Management 63, no. 1: 104262. https://doi.org/10.1016/j.ipm.2025.104262.

Babaei Giglou, Hamed, Jennifer D'Souza, and Sören Auer. 2023. "LLMs4OL: Large Language Models for Ontology Learning." In The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part I, edited by Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, 408–427. Lecture Notes in Computer Science 14265. Cham: Springer. https://doi.org/10.1007/978-3-031-47240-4_22.

Cao, Zhihan, Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. 2025. "A Comprehensive Evaluation of Semantic Relation Knowledge of Pretrained Language Models and Humans." Language Resources and Evaluation 59: 3799–3849. https://doi.org/10.1007/s10579-025-09858-9.

Chen, Banghao, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. "Unleashing the Potential of Prompt Engineering for Large Language Models." Patterns 6, no. 6: 101260. https://doi.org/10.1016/j.patter.2025.101260.

Chen, Shu-Jiun. 2025. "Advancing Thesaurus Construction With Generative AI: A Structured Approach to Synonym Identification and Scope Note Development." Knowledge Organization 52, no. 5: 46865. https://doi.org/10.31083/KO46865.

He, Yuan, Jiaoyan Chen, Hang Dong, and Ian Horrocks. 2023. "Exploring Large Language Models for Ontology Alignment." arXiv:2309.07172. https://doi.org/10.48550/arXiv.2309.07172.

International Organization for Standardization. 2011. ISO 25964-1:2011. Information and Documentation — Thesauri and Interoperability with Other Vocabularies — Part 1: Thesauri for Information Retrieval. Geneva: ISO.

International Organization for Standardization. 2013. ISO 25964-2:2013. Information and Documentation — Thesauri and Interoperability with Other Vocabularies — Part 2: Interoperability with Other Vocabularies. Geneva: ISO.

Kempf, Andreas Oskar, and Joachim Neubert. 2021. "STW Thesaurus for Economics." In ISKO Encyclopedia of Knowledge Organization, edited by Birger Hjørland and Claudio Gnoli. https://www.isko.org/cyclo/stw.

Li, Jiayi, Daniel Garijo, and María Poveda-Villalón. 2026. "Large Language Models for Ontology Engineering: A Systematic Literature Review." Semantic Web 17, no. 4: 1–45. https://doi.org/10.1177/22104968261465514.

Proietti, Mattia, and Alessandro Lenci. 2025. "The Quasi-Semantic Competence of LLMs: A Case Study on the Part-Whole Relation." arXiv:2504.02395. https://doi.org/10.48550/arXiv.2504.02395.

Roussinov, Dmitri, Serge Sharoff, and Nadezhda Puchnina. 2023. "Fine-Tuning Language Models to Recognize Semantic Relations." Language Resources and Evaluation 57: 1463–1486. https://doi.org/10.1007/s10579-023-09677-w.

Shimizu, Cogan, and Pascal Hitzler. 2025. “Accelerating Knowledge Graph and Ontology Engineering with Large Language Models.” Journal of Web Semantics 85: 100862. https://doi.org/10.1016/j.websem.2025.100862.

Appendix A. Optimized Prompt Strategy within Six-Component Prompt Framework

We applied the following optimized prompt strategy within the six-component prompt framework:

Table A1. Applied Six-Component Prompt Framework

Dimension	Component	Content
Task Design	Task Role (assigns AI a specialized role)	You are the editor-in-chief of the internationally used, bilingual STW Thesaurus for Economics.
	Task Context (provides relevant background)	In the English version of the STW Thesaurus for Economics, you deal with concepts, resp. descriptor records, from the sub-thesauri Economics, Business economics, Economic sectors, Commodities and Related subject areas. For the English version, you want to specify the relationship between the English preferred term, resp. descriptor, and its corresponding English non-preferred terms, resp. non-descriptors. You will receive the following three parameters as input data: - preferred term resp. descriptor - non-preferred term resp. non-descriptor - context given in the form of one or more STW subject categories
	Task Goals (a-c) (defines the explicit objective)	a) define each descriptor in one English sentence, b) evaluate the relationship between descriptor and corresponding non-descriptor according to a pre-defined rating scale, c) explain each rating with one or more arguments in one English sentence.
Guidance Design	Instruction Rules (specifies analytical methods)	Instructions: (a) Definition/Scope note generation: Define each descriptor in one English sentence. (b) Relationship assessment between descriptor and non-descriptor (see Table 1) (c) Explanation: Explain each rating with one or more arguments in one

Dimension	Component	Content
		English sentence.
	Instruction Examples (provides structured examples of expected outputs)	Example output: “preferred term”: “...” “definition”: “...” “non_preferred term”: “...” “scale value”: “...” “explanation”: “...”

	Instruction Format (standardizes output presentation format)	Present the results in JSON format as follows:
		{
		"descriptor": "...",
		"definition": "...",
		"evaluation": [
		{
		"non_descriptor": "...",
		"scale value": 0-4,
		"explanation": "..."
		},
		...
		]
		}

Appendix B. Evaluation Management System Interface

The LLM-based and intellectual evaluations are presented in a dedicated interface. Users can select one of the evaluated LLMs from a selection menu. After the intellectual assessment has been entered, the interface displays the level of agreement between the selected LLM and the expert evaluation. In the search box below, descriptors can be searched. On the left, the preferred term is shown together with an LLM-generated definition. Indicators in the upper right corner display the total number of non-descriptors and how many have already received an intellectual assessment. On the right, the LLM-generated definition appears below the preferred term, followed by the number of non-descriptor pairs. For each non-descriptor, the interface displays the LLM-based classification and the intellectual assessment side by side, along with the LLM's explanation. At the bottom of each entry, users can record the intellectual evaluation and add a note.

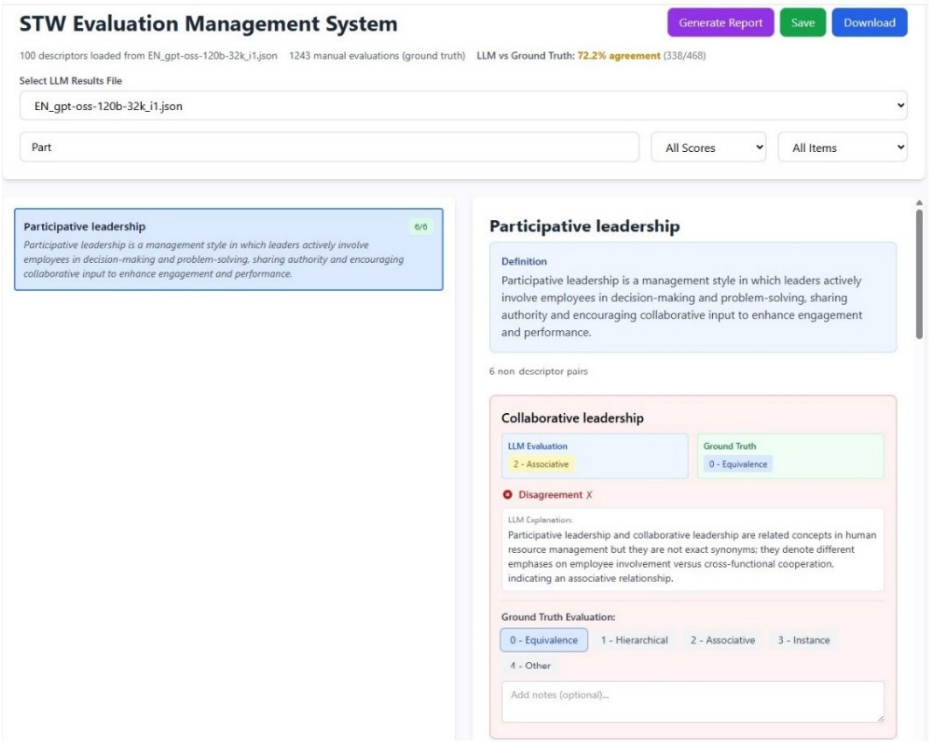

Figure B1. Display of the evaluation management system interface.

Comparative Performance Metrics

LLM Model	Accuracy	Precision (macro)	Recall (macro)	F1-Score (macro)
EN - ensemble details	**75.64%**	**42.87%**	**39.29%**	**40.63%**
DE - ensemble details	74.90%	41.43%	43.13%	41.92%
EN - gpt-oss-120b_i1 details	72.22%	41.89%	37.66%	39.58%
EN - qwen3-32b_i2 details	70.51%	40.68%	41.08%	40.56%
DE - gpt-oss-20b_i2 details	70.22%	47.17%	50.55%	44.11%
DE - gpt-oss-20b_i1 details	69.81%	48.54%	66.97%	49.23%
DE - gpt-oss-120b_i1 details	69.57%	46.04%	54.19%	47.29%
DE - gpt-oss-120b_i2 details	69.30%	44.76%	53.43%	44.83%
EN - gpt-oss-20b_i1 details	69.02%	41.49%	36.91%	39.05%
EN - deepseek-r1-70b_i2 details	68.19%	38.63%	36.50%	37.36%
EN - gpt-oss-20b_i2 details	67.24%	38.55%	35.13%	36.75%
EN - llama3.3-70b details	67.02%	36.59%	37.23%	35.28%
DE - qwen3-32b_i2 details	63.32%	39.98%	52.73%	40.23%
EN - deepseek-r1-8b_i2 details	61.52%	32.09%	25.13%	27.35%
DE - deepseek-r1-70b_i2 details	59.90%	41.22%	53.47%	42.78%
DE - llama3.3-70b_i1 details	58.69%	47.35%	47.49%	45.34%
EN - mistral-8b_i1 details	56.96%	29.30%	30.84%	28.85%
EN - deepseek-r1-8b_i1 details	50.54%	27.08%	24.49%	25.38%
EN - apertus-8b_i1 details	45.80%	35.36%	19.24%	21.27%
EN - llama3-8b_i1 details	44.21%	28.18%	28.10%	23.23%

Figure B2. Display of the LLM evaluation report – by clicking on “details” the user can find all comparative performance metrics for each model.